ZigZag: A new approach to adaptive online learning

Cornell University

Dylan Foster, Alexander Rakhlin, and Karthik Sridharan

djfoster@cs.cornell.edu, rakhlin@wharton.upenn.edu, sridharan@cs.cornell.edu

Penn University of Pennsylvania

Overview

Online supervised learning

Protocol:
For $t = 1$ to n:

Receive input instance $x_t \in \mathcal{X}$.

Learner picks (possibly randomized) prediction $\hat{y}_t \in \mathbb{R}$.

Receive outcome $y_t \in \mathcal{Y}$.

Learner suffers loss $\ell(\hat{y}_t, y_t)$ (convex, 1-Lipschitz wrt $\hat{y}$).

End

Performance measure: regret w.r.t. hypothesis class $\mathcal{F}$:

$$\mathbf{Reg}_n(x_{1:n}, y_{1:n}) = \sum_{t=1}^n \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^n \ell(f(x_t), y_t)$$

Goal: Adapt to data instances.

Our notion of adaptivity:
Empirical Rademacher Complexity

$$\widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}) = \mathbb{E}_{\epsilon} \sup_{f\in\mathcal{F}} \sum_{t=1}^n \epsilon_t f(x_t)$$

Questions we study:

- When does there exist a strategy $(\hat{y}_t)$ such that

$$\sum_{t=1}^n \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^n \ell(f(x_t), y_t) \le \mathbf{D}(\mathcal{F}, n) \cdot \widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n})$$

  for *every sequence* $x_{1:n}, y_{1:n}$?
- What is the best constant $\mathbf{D}(\mathcal{F}, n)$?
- When can the strategy $(\hat{y}_t)$ be efficiently computed?

Why $\widehat{\mathbf{Rad}}_{\mathcal{F}}$?

Lemma: Sequence Optimality

If there is a strategy $(\hat{y}_t)$ and function $\mathcal{B}$ with

$$\sum_{t=1}^n \ell_{\mathrm{abs}}(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^n \ell_{\mathrm{abs}}(f(x_t), y_t) \le \mathcal{B}(x_{1:n}),$$

then $\widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}) \le \mathcal{B}(x_{1:n})$ *pointwise.*

In particular, gives optimal online-to-batch.

Linear Classes

Data space (where x_ts live): $\mathcal{X} = (\mathfrak{B}, \|\cdot\|)$.
Natural choice for function class:

$$\mathcal{F} = \{x \mapsto \langle w, x\rangle \mid w \in \mathfrak{B}^\star, \|w\|_\star \le 1\}$$

In this case,

$$\widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}) = \mathbb{E}_{\epsilon} \sup_{\|w\|_\star\le 1} \sum_{t=1}^n \epsilon_t \langle w, x_t\rangle = \mathbb{E}_{\epsilon} \left\| \sum_{t=1}^n \epsilon_t x_t \right\|.$$

Example: Matrix Prediction

Matrix prediction ($\mathfrak{B} = \mathbb{R}^{d\times d}$):

- $x_t = e_{i_t} e_{j_t}^\top$ (entry we want to predict).
- $\mathcal{F} = \left\{ f \in \mathbb{R}^{d\times d} \mid \|f\|_{\mathrm{trace}} \le r\sqrt{d} \right\}$ (contains all rank-r matrices with ℓ_∞ bounded entries).

$$\widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}) \le \begin{cases} \sqrt{rd}\cdot\sqrt{n} & \text{entries IID} \\ \sqrt{rd}\cdot\sqrt{n} & \text{worst case} \\ \sqrt{rd}\cdot\sqrt{(\#\text{ row rep.}) \vee (\#\text{ col rep.})} & \text{general case} \end{cases}$$

[Foygel-Srebro'11]

Main results

New framework for developing efficient and adaptive online learning algorithms

- New algorithms: ℓ_p AdaGrad generalization and more.
- Beyond mirror descent!

Tight characterization of $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ achievability based on decoupling inequalities for martingales.

- Exploit characterization algorithmically.
- New rates for low rank experts, empirical covering, etc.

The algorithm: ZigZag

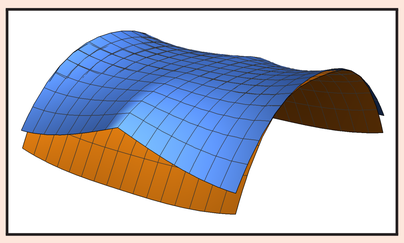
ZigZag: Recipe for achieving $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ for linear classes

ZigZag algorithm is parameterized by $\mathbf{U} : \mathfrak{B} \times \mathfrak{B} \to \mathbb{R}$:

Definition: Burkholder Function

$\mathbf{U}$ is a Burkholder function for $(\|\cdot\|, p)$ if there is constant $\mathbf{C}_p$ such that

1. $\mathbf{U}(x, x') \ge \|x\|^p - \mathbf{C}_p^p \|x'\|^p$, for some $p \ge 1$.
2. $G(z) := \mathbf{U}(x + z, x' + \epsilon z)$ is concave for all $x, x' \in \mathfrak{B}$ and $\epsilon \in \{\pm 1\}$.
   ("zig-zag concavity")



Algorithm: ZigZag

Input: Burkholder fn. $\mathbf{U}$ for $(\|\cdot\|, p)$, learning rate η.

For time $t = 1, \dots, n$:

- Define $G_t(\alpha) = \mathbb{E}_{\sigma_t\in\{\pm1\}} \frac{\eta}{p} \mathbf{U}_p\left(\sum_{s=1}^{t-1} \ell'_s x_s + \alpha x_t, \sum_{t=1}^{t-1} \epsilon_s \ell'_s x_s + \sigma_t \alpha x_t\right)$.
  (Abbreviating $\ell'(\hat{y}, y_t)$ to ℓ'_t)
- Predict $\hat{y}_t = -G'_t(0)$.
- Draw independent Rademacher $\epsilon_t \in \{\pm 1\}$.

*for $p \ne 1$, additional learning rate tuning on top of this scheme

Theorem: Regret of ZigZag

Given a Burkholder function $\mathbf{U}$ as above, **ZigZag** guarantees

$$\sum_{t=1}^n \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^n \ell(f(x_t), y_t) \le \mathbf{C}_p \cdot \widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}).$$

Proof sketch (using $\eta = p = 1$ to simplify)

Our goal is to show the following quantity is at most zero:

$$\sum_{t=1}^n \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^n \ell(f(x_t), y_t) - \mathbf{D} \cdot \widehat{\mathbf{Rad}}_{\mathcal{F}}(x_{1:n}).$$

Linearizing and expanding out $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ (+contraction), we have

$$\begin{aligned}
&\le \sum_{t=1}^n \hat{y}_t \ell'_t + \left\| \sum_{t=1}^n \ell'_t x_t \right\| - \mathbf{D} \cdot \mathbb{E}_{\epsilon} \left\| \sum_{t=1}^n \epsilon_t \ell'_t x_t \right\| \\
&\le \sum_{t=1}^n \hat{y}_t \ell'_t + \mathbb{E}_{\epsilon} \mathbf{U}\left(\sum_{t=1}^n \ell'_t x_t, \sum_{t=1}^n \epsilon_t \ell'_t x_t \right) \\
&= \sum_{t=1}^{n-1} \hat{y}_t \ell'_t + \hat{y}_n \ell'_n + \mathbb{E}_{\epsilon} G_n(\ell'_n) \\
&= \sum_{t=1}^{n-1} \hat{y}_t \ell'_t + \mathbb{E}_{\epsilon}[-G'_n(0)\ell'_n + G_n(\ell'_n)] \\
&\le \sum_{t=1}^{n-1} \hat{y}_t \ell'_t + \mathbb{E}_{\epsilon} G_{n-1}(\ell'_{n-1}) \quad \text{Using zig-zag concavity of } \mathbf{U}. \\
&\le \ldots\text{proceed inductively.}
\end{aligned}$$

ZigZag example

Example: ℓ_p norms

Zig-zag function for any ℓ_p:

$$\mathbf{U}(x, x') = C_p \sum_{i=1}^d (|x_i| - c_p|x'_i|)(|x_i| + |x'_i|)^{p-1} \quad \text{[Burkholder'84]}$$

ZigZag algorithm with $\mathbf{U}$ gives

$$\sum_{t=1}^n \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^n \ell(f(x_t), y_t) \le \left(p \vee \frac{p}{p-1}\right)^2 \cdot \mathbb{E} \left\| \sum_{t=1}^n x_t \right\|_{\ell_p}$$

Recovers AdaGrad:

$$\mathbb{E}\left\| \sum_{t=1}^n x_t \right\|_{\ell_2} \asymp \text{Full matrix AdaGrad bound [Duchi-Hazan-Singer'11]}$$

$$\mathbb{E}\left\| \sum_{t=1}^n x_t \right\|_{\ell_1} \asymp \text{Diagonal AdaGrad bound*}$$

Gives new AdaGrad generalization to general ℓ_p norms!

*technical detail: Use $p = 1 + 1/\log(d)$ in ℓ_p alg.

Example above: $O(d)$ time/update.
Most examples: Time $O(\dim)$, e.g. $O(d^2)$ for matrix w/ group norm.

$\mathbf{U}$ functions and decoupling inequalities

Donald Burkholder (1984): **$\mathbf{U}$ functions are equivalent to decoupling inequalities for martingales!**

Unconditional Martingale Difference (UMD) property

$(\mathfrak{B}, \|\cdot\|)$ is UMD if $\exists\, p \ge 1$ and constant $\mathbf{C}_p$ such that for all MDS $(Z_t)_{t\le n} \in \mathfrak{B}$ and all sign sequences $(\epsilon_t)_{t\le n} \in \{\pm 1\}$,

$$\mathbb{E}\left\| \sum_{t=1}^n \epsilon_t Z_t \right\|^p \le \mathbf{C}_p^p\, \mathbb{E}\left\| \sum_{t=1}^n Z_t \right\|^p.$$

Theorem (Burkholder '84)

UMD holds for $(\mathfrak{B}, \|\cdot\|)$ with constant $\mathbf{C}_p$ iff there exists Burkholder fn. $\mathbf{U}$ for $\|\cdot\|$, same constant.

Necessary vs. sufficient conditions

Achievability of $\widehat{\mathbf{Rad}}_{\mathcal{F}}$ implies a UMD-type inequality even for general function classes!

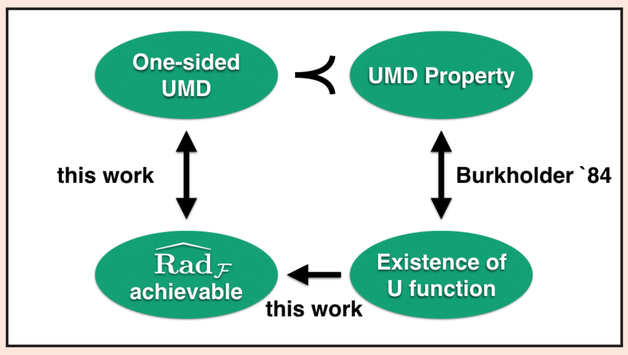
UMD property for general function classes

$\mathcal{F}$ satisfies generalized (one-sided) UMD if:
For any $\mathcal{X}$-valued tree $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ where each $\mathbf{x}_t : \{\pm 1\}^{t-1} \to \mathcal{X}$, $\exists \mathbf{C}_p$ such that

$$\mathbb{E}_{\epsilon}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^n \epsilon_t f(\mathbf{x}_t(\epsilon_{1:t-1})) \right]^p \le \mathbf{C}_p^p\, \mathbb{E}_{\epsilon,\epsilon'}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^n \epsilon'_t f(\mathbf{x}_t(\epsilon_{1:t-1})) \right]^p.$$

$\epsilon = (\epsilon_1, \dots, \epsilon_n)$ and $\epsilon' = (\epsilon'_1, \dots, \epsilon'_n)$ are independent Rademacher random variables.

Characterization:



Theorem: Equivalence for General Classes

Let $\mathcal{F} \subset [-1, 1]^{\mathcal{X}}$ be any class of predictors. The following statements are equivalent:

1. There exists a learning algorithm and constant B such that the following regret bound against any adversary holds:

$$\sum_{t=1}^n \ell_{\mathrm{abs}}(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \sum_{t=1}^n \ell_{\mathrm{abs}}(f(x_t), y_t) \le B\, \mathbb{E}_{\epsilon} \sup_{f\in\mathcal{F}} \sum_{t=1}^n \epsilon_t f(x_t)$$

2. There exists constant $\mathbf{C}_1$ for which generalized one-sided UMD holds with $p = 1$.